先验分布、后验分布及贝叶斯学习

彭轩宇 1900013014

2019 年 10 月 26 日

概率

- 条件概率：$P(A|B) = \frac{P(AB)}{P(B)}$。
- 联合分布：$P(x, y)$，当 X, Y 独立时有 $P(x, y) = P(x)P(y)$。
- 贝叶斯公式：

$$P(B_i|A) = \frac{P(AB_i)}{P(A)} = \frac{P(B_i)P(A|B_i)}{\sum_j P(B_j)P(A|B_j)}$$

- 这里的贝叶斯公式虽然看起只是平凡的数学推导，但其之所以著名在于其现实乃至哲理的意义上。

估计

估计参数

给定样本 $X_1, ..., X_n$，样本是从某一分布中取出，估计这个分布的参数 θ。

- 例如，某厂家生产的电子产品寿命服从指数分布 $P(x|\theta) = \theta e^{-\theta x}$，现在随机抽取了一些产品检测其寿命，即得到样本集 $\{X_n\}$，估计参数 θ。
- 当然 θ 指的可以是一个参数集，比如说正态分布就有两个参数 μ, σ。
- 有些情况还需要做一个预测，若再给定一个 X_{new}，估计 $P(X_{new})$。

似然

- 在给定样本集 $\{X_n\}$ 的，样本之间可以视为互相独立，那么 $P(x|\theta) = \prod_{i=1}^{n} P(x_i|\theta)$。
- 如果把 $P(x|\theta)$ 看作是 θ 的函数时，我们称为**“似然函数”**，记作 $f(x|\theta)$。
- “似然”这个名称的意义，反映了在观测结果 $\{X_n\}$ 已知的条件下，如果参数取 θ 的“相似程度”，像是一个由果推因。在频率学派的观点下，θ 并非事件或者随机变量，无概率可言，所以用似然 (likelihood) 这个词。

极大似然估计 (MLE)

MLE

$$\theta^* = \arg\max_{\theta} P(x|\theta)$$

- MLE 就是最大化似然函数，比较直观，θ^* 就是估计结果。
- 如果再给定一个 X_{new}，直接用 $P(X_{new}|\theta^*)$ 作为预测结果。

从先验分布到后验分布

- 而贝叶斯派认为参数 θ 也是一个随机变量，且我们对 θ 已经有一定的认知，也就是把 θ 认为是一个分布 $\pi(\theta)$，即**先验分布**。
- 对于上文中的例子，我们已经对电子产品的质量有一些主观认知。
- 根据贝叶斯公式，就有

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\theta)\pi(\theta)\,d\theta}$$

- 其中 $\pi(\theta|x)$ 就是**后验分布**，后验分布综合了样本信息和先验分布，也就是用样本更新了先前对 θ 的认知。

极大后验估计 (MAP)

MAP

$$\theta^* = \arg\max_{\theta} \pi(\theta|x)$$

- 就是把最大化函数变成了后验分布。
- 回顾一下朴素贝叶斯分类器：相当于我们有只一个样本 $\mathbf{x}$（即待分类的文章向量），要估计它的类别 θ^*。
- 已经分好类的文章相当于是先验知识，以及 $\pi(\theta|x)$ 的分母是一个与 θ 无关的东西。
- 又因为 $\mathbf{x}$ 的每一维都是独立的，似然函数也很好计算。MAP 最大化 $\pi(\theta|\mathbf{x})$ 就可以找到 θ^*

贝叶斯估计

- 不断地试验获取样本，并且更新之前的认知，即不断地用后验分布作为新的先验分布，就是一个贝叶斯学习的过程。
- 同时，做预测的时候，贝叶斯估计不再用某一个参数来做预测而是通过整个后验分布做预测。

贝叶斯估计

$$P(X_{new}) = \int \pi(\theta|x) P(X_{new}|\theta)\, d\theta$$

- 相当于是把后验分布当做 θ 在预测中的权重分布。
- 但是如果要做预测的话，后验分布中的分母 $\int_\Theta f(x|\theta)\pi(\theta)\, d\theta$ 就不能丢弃，且成为算法的了算法的瓶颈。于是我们引入**共轭分布**作为一种解决方案。

共轭分布

- 考虑抛硬币的例子（伯努利试验），要预测它抛出正面的概率 p。
- 假设在样本数据中抛出了 m_h 次正面，m_t 次反面，$f(x|\theta) = \theta^{m_h}(1-\theta)^{m_t}$。
- 因为随着样本的增多，最初的先验知识所占比重越来越小，所以第一次先验分布其实是可以任意选取的。
- 我们能否找到一种分布作为先验，使得经过贝叶斯估计后，后验也是这种分布？对于这个例子，Beta 分布就是合适的。

共轭分布

Beta 分布

$$Beta(\theta|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

其中

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

- Γ 函数是对阶乘的拓展，这个例子中只需把 B 函数理解成组合数，$Beta$ 分布理解成二项分布就好了。

共轭分布

- 假设在以前的试验中有 α_h 次正面，α_t 次反面，选取先验为 $Beta(\alpha_h, \alpha_t) = \theta^{\alpha_h - 1}(1-\theta)^{\alpha_t - 1}$，即

$$\pi(\theta) \propto \theta^{\alpha_h - 1}(1-\theta)^{\alpha_t - 1}$$

- 进行贝叶斯估计，乘上似然

$$\pi(\theta|x) \propto \theta^{m_h + \alpha_h - 1}(1-\theta)^{m_t + \alpha_t - 1}$$

- 标准化后发现后验仍然是 Beta 分布

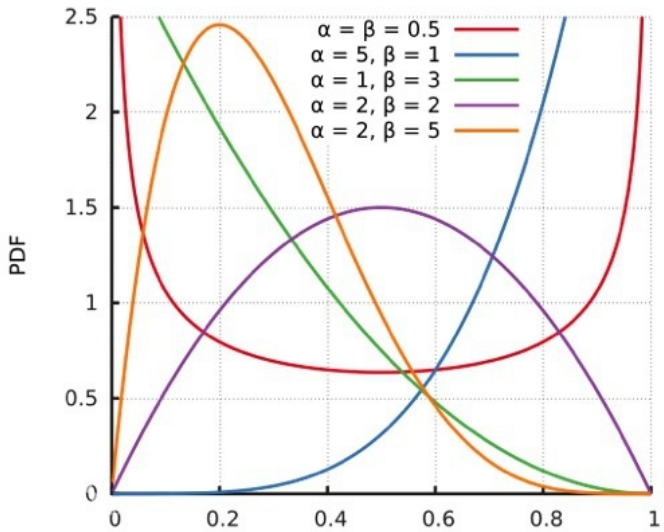
$$\pi(\theta|x) = \frac{1}{B(m_h + \alpha_h, m_t + \alpha_t)}\theta^{m_h + \alpha_h - 1}(1-\theta)^{m_t + \alpha_t - 1}$$

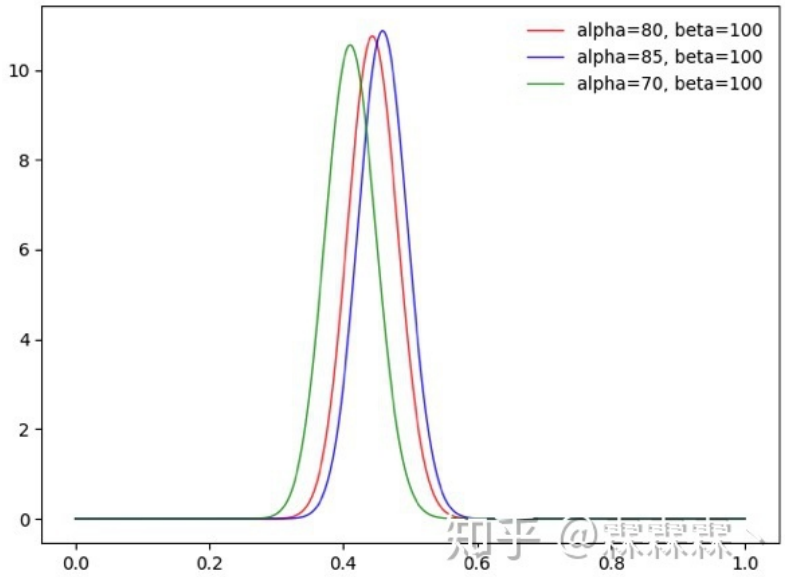
$$= Beta(m_h + \alpha_h, m_t + \alpha_t)$$

共轭分布

- 于是我们并没有真正的进行复杂的积分计算，每次学习仅仅是调整 Beta 分布的两个参数即可。
- 总结一下，就是对于似然服从伯努利分布，先验服从 Beta 分布，后验也服从 Beta 分布，于是称伯努利分布的**共轭先验**是 Beta 分布。
- 还有很多例子，泊松分布的共轭先验是伽马分布，正态分布的共轭先验是正态分布等等。
- 当然除了利用共轭分布，还有其他方法进行贝叶斯估计。

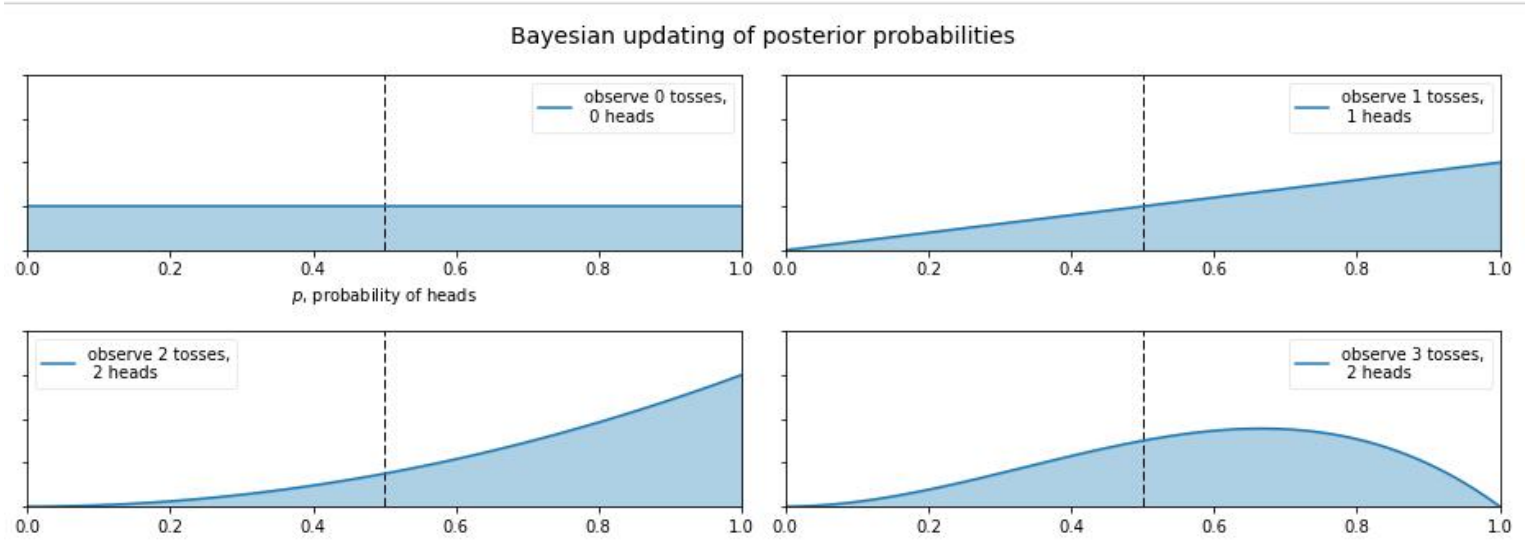
共轭分布





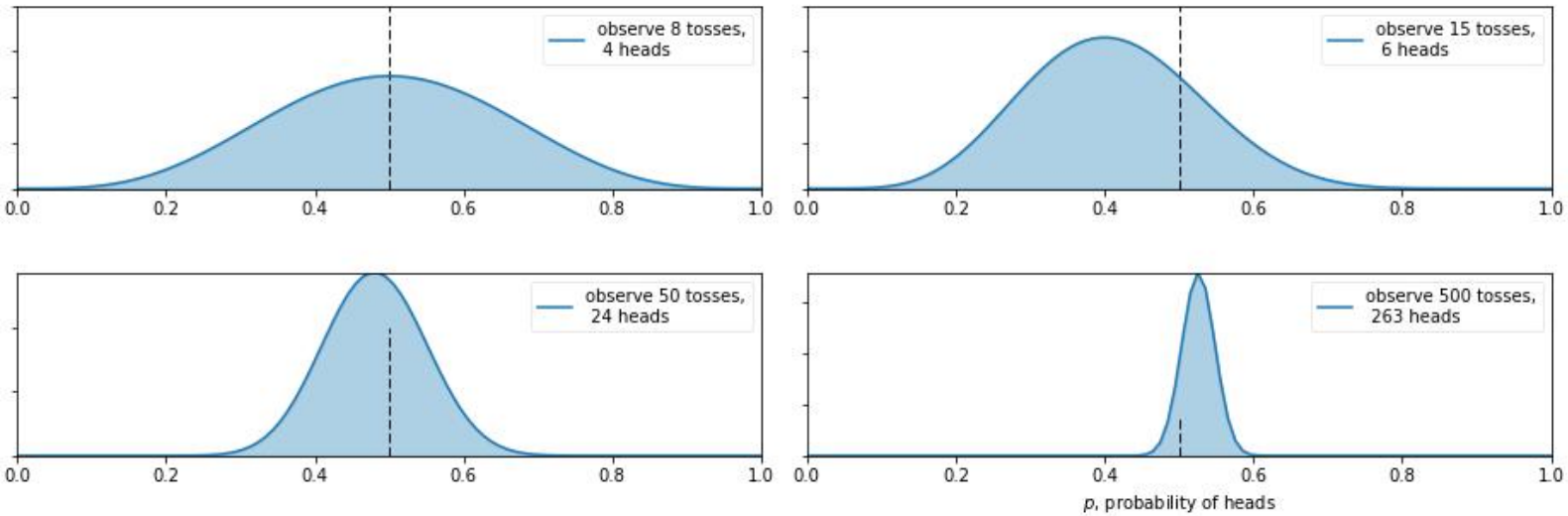
- 注意到 Beta 分布形状多变，可以呈现不同的形式。

效果



- 每次都是把上一张图作为先验，然后得出当前这张图。这是抛了 3 次的情况。
- 可能会有一些疑问，觉得这和频率学派在大量实验后得出概率没什么区别。事实上，假如前两次都抛出“正”，频率学派会预测抛出正的概率 $p = 1$，而贝叶斯估计的出的是 p 的一个分布（左下图），如果预测下一次为正的概率，并不是 1。

效果



- 可以看到最终在 0.5 处出现了一个陡峭的峰值。说明选取的这个硬币是均匀的。

参考

- 《概率论与数理统计》陈希孺，中国科学技术大学出版社
- https://wso2.com/blog/research/part-one-introduction-to-bayesian-learning
- http://www.xuyankun.cn/2017/05/13/bayes/
- https://www.cnblogs.com/Luv-GEM/p/10638480.html
- https://zhuanlan.zhihu.com/p/72508229